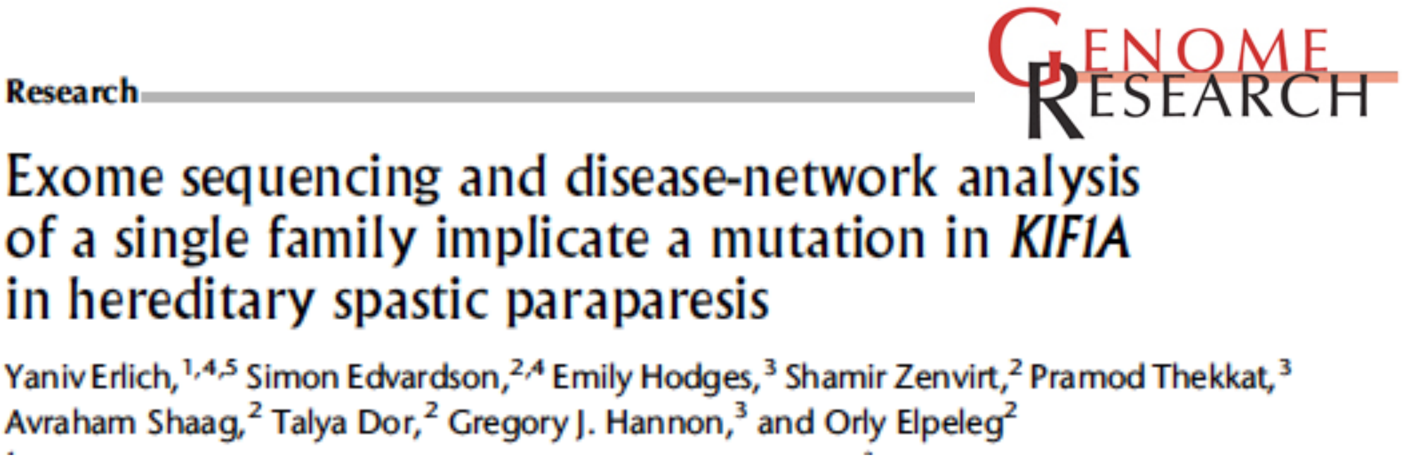
Genome Hacking

Yaniv Erlich
Whitehead Institute for Biomedical Research

Twitter: @erlichya

Public data is important for genetic studies

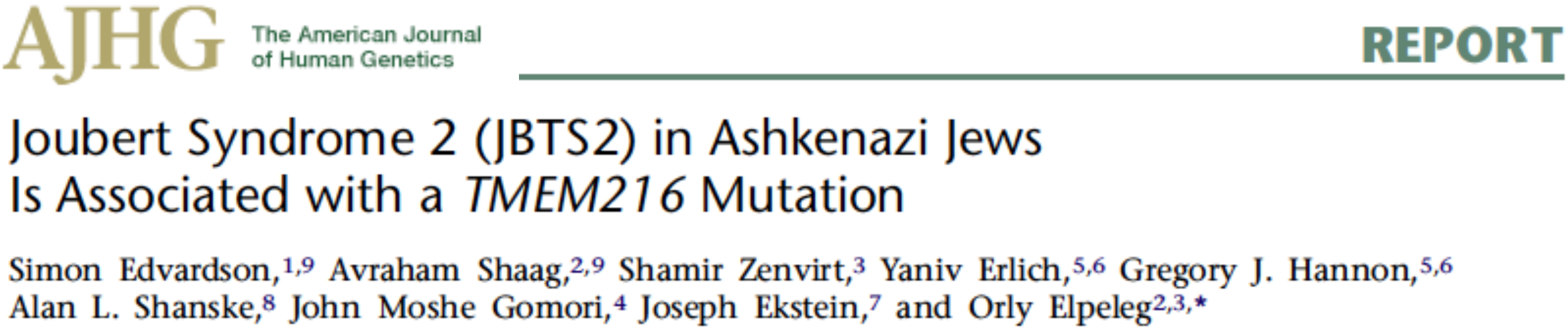


Research

GENOME RESEARCH

Exome sequencing and disease-network analysis of a single family implicate a mutation in *KIF1A* in hereditary spastic paraparesis

Yaniv Erlich,[1,4,5] Simon Edvardson,[2,4] Emily Hodges,[3] Shamir Zenvirt,[2] Pramod Thekkat,[3] Avraham Shaag,[2] Talya Dor,[2] Gregory J. Hannon,[3] and Orly Elpeleg[2]



AJHG The American Journal of Human Genetics

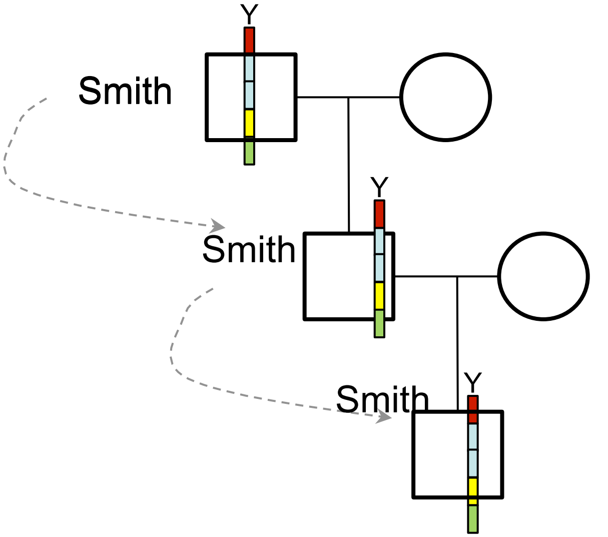
REPORT

Joubert Syndrome 2 (JBTS2) in Ashkenazi Jews Is Associated with a *TMEM216* Mutation

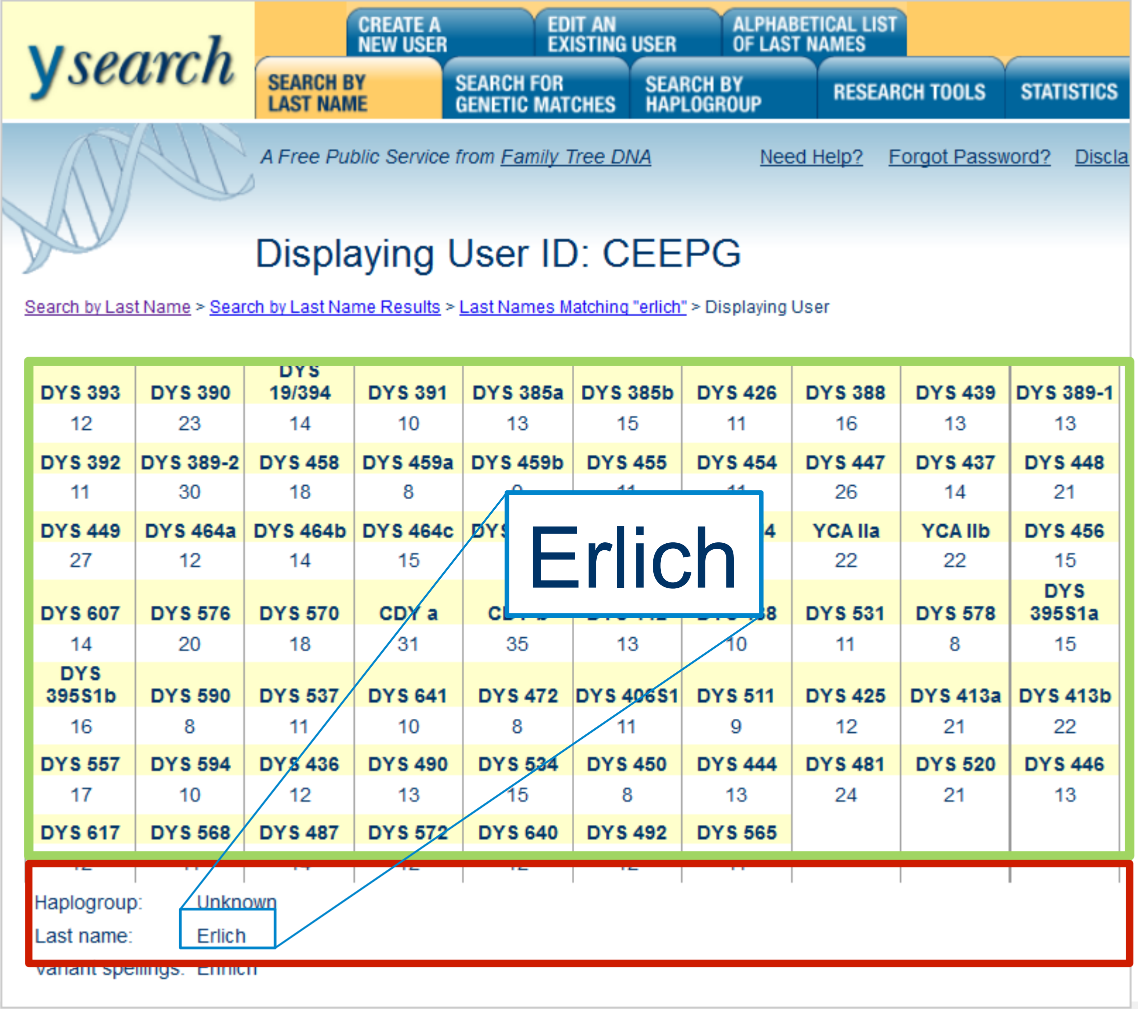
Simon Edvardson,[1,9] Avraham Shaag,[2,9] Shamir Zenvirt,[3] Yaniv Erlich,[5,6] Gregory J. Hannon,[5,6] Alan L. Shanske,[8] John Moshe Gomori,[4] Joseph Ekstein,[7] and Orly Elpeleg[2,3,*]

To make this endeavor sustainable, we must proactively map risks

Co-segregation between Y-chr and surnames



www.ysearch.org:



Displaying User ID: CEEPG

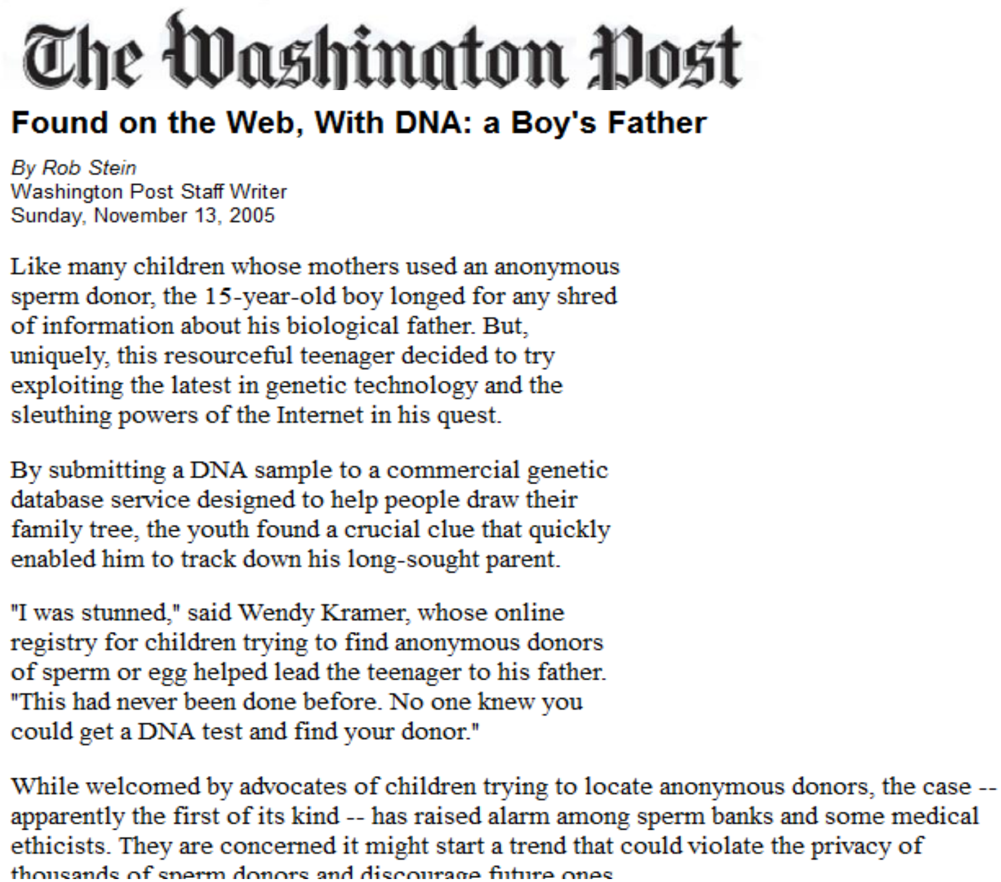
Search by Last Name > Search by Last Name Results > Last Names Matching "erlich" > Displaying User

DYS 393	DYS 390	DYS 19/394	DYS 391	DYS 385a	DYS 385b	DYS 426	DYS 388	DYS 439	DYS 389-1
12	23	14	10	13	15	11	16	13	13
DYS 392	DYS 389-2	DYS 458	DYS 459a	DYS 459b	DYS 455	DYS 454	DYS 447	DYS 437	DYS 448
11	30	18	8	9	11	11	26	14	21
DYS 449	DYS 464a	DYS 464b	DYS 464c	[illegible]	[illegible]	[illegible]	YCA IIa	YCA IIb	DYS 456
27	12	14	15	[illegible]	[illegible]	[illegible]	22	22	15
DYS 607	DYS 576	DYS 570	CDY a	CDY b	[illegible]	[illegible]	DYS 531	DYS 578	DYS 395S1a
14	20	18	31	35	13	10	11	8	15
DYS 395S1b	DYS 590	DYS 537	DYS 641	DYS 472	DYS 406S1	DYS 511	DYS 425	DYS 413a	DYS 413b
16	8	11	10	8	11	9	12	21	22
DYS 557	DYS 594	DYS 436	DYS 490	DYS 534	DYS 450	DYS 444	DYS 481	DYS 520	DYS 446
17	10	12	13	15	8	13	24	21	13
DYS 617	DYS 568	DYS 487	DYS 572	DYS 640	DYS 492	DYS 565			
12	11	14	12	12	12	11			

Haplogroup: Unknown
Last name: Erlich
Variant spellings: Ehrlich

Erlich

Exploiting genetic genealogy databases



The Washington Post

Found on the Web, With DNA: a Boy's Father

By Rob Stein
Washington Post Staff Writer
Sunday, November 13, 2005

Like many children whose mothers used an anonymous sperm donor, the 15-year-old boy longed for any shred of information about his biological father. But, uniquely, this resourceful teenager decided to try exploiting the latest in genetic technology and the sleuthing powers of the Internet in his quest.

By submitting a DNA sample to a commercial genetic database service designed to help people draw their family tree, the youth found a crucial clue that quickly enabled him to track down his long-sought parent.

"I was stunned," said Wendy Kramer, whose online registry for children trying to find anonymous donors of sperm or egg helped lead the teenager to his father. "This had never been done before. No one knew you could get a DNA test and find your donor."

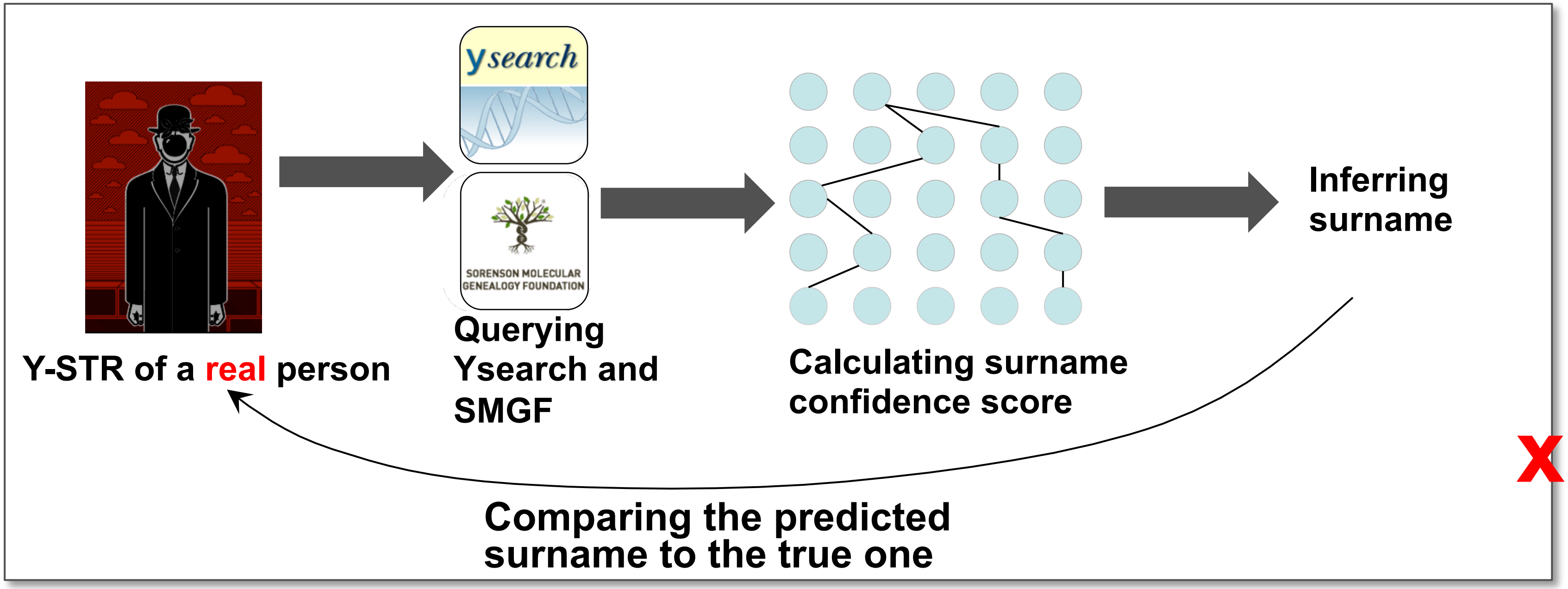
While welcomed by advocates of children trying to locate anonymous donors, the case -- apparently the first of its kind -- has raised alarm among sperm banks and some medical ethicists. They are concerned it might start a trend that could violate the privacy of thousands of sperm donors and discourage future ones.

An anecdote?

The main idea – a systematic study

Can we recover the identity of anonymous sequencing datasets using public resources?

Empirical test: what is the probability to recover a surname?



Y-STR of a real person

Querying Ysearch and SMGF

Calculating surname confidence score

Inferring surname

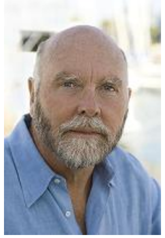
Comparing the predicted surname to the true one

x900

Expectation for US Caucasian males from middle and upper class:

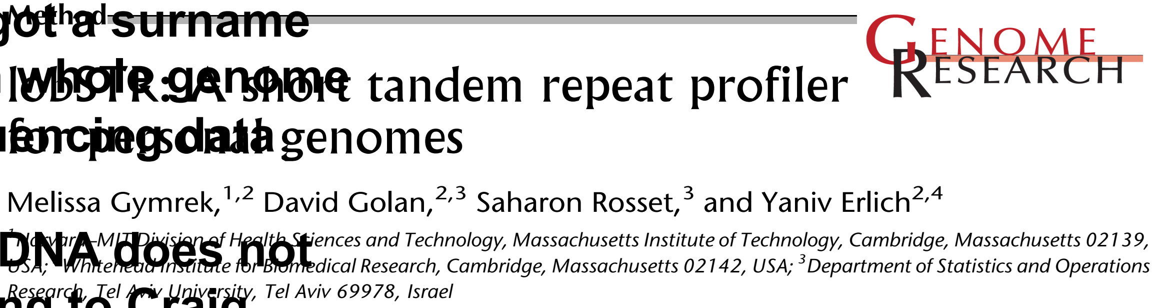
12% Successful recoveries

The Venter case



lobSTR

- **We got a surname from whole genome sequencing data**
- **The DNA does not belong to Craig Venter**



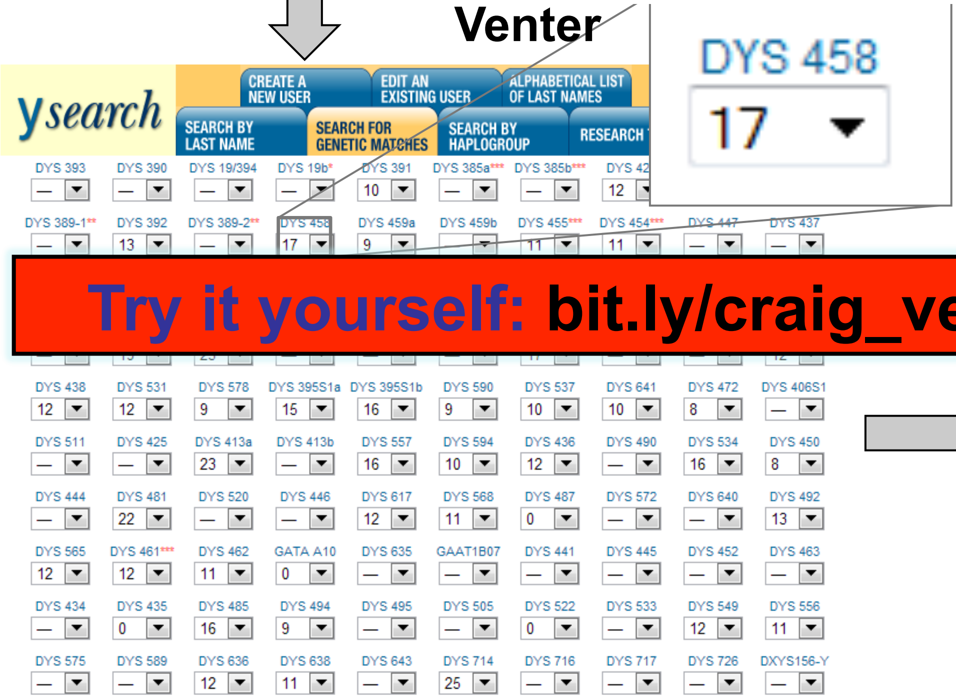
Method

GENOME RESEARCH

lobSTR: A short tandem repeat profiler for personal genomes

Melissa Gymrek,[1,2] David Golan,[2,3] Saharon Rosset,[3] and Yaniv Erlich[2,4]

[1]Harvard–MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; [2]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA; [3]Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel

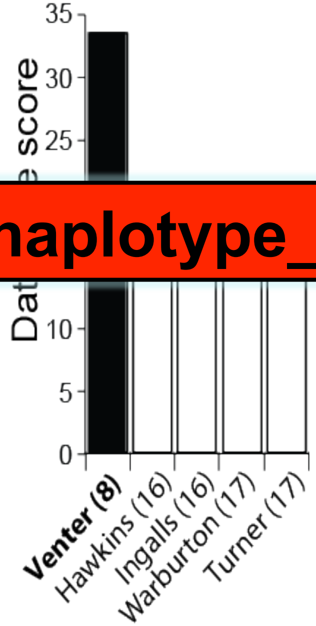


DYS 458

17

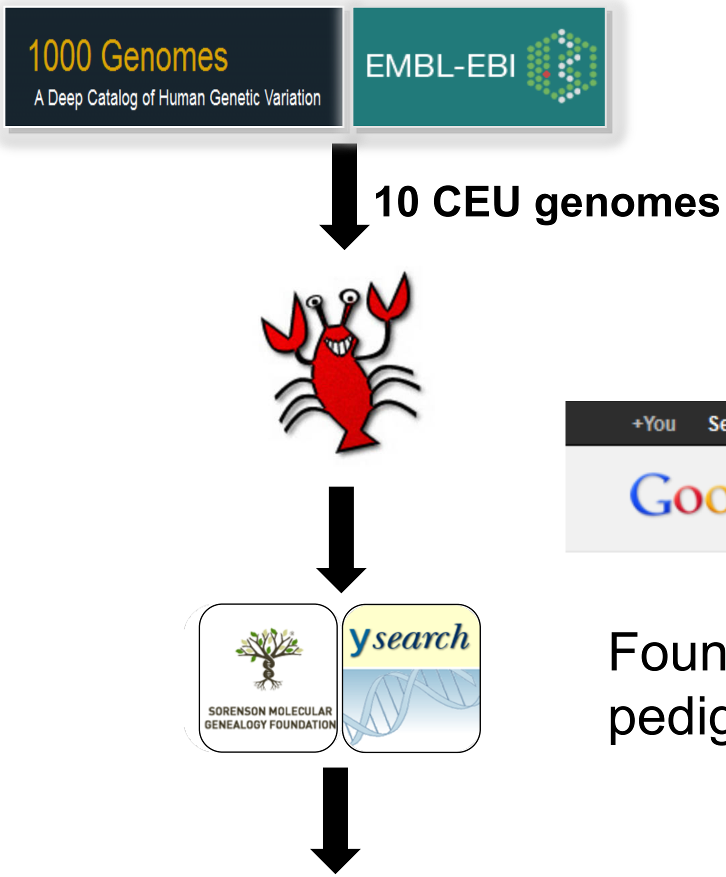
DYS 393	DYS 390	DYS 19/394	DYS 19b*	DYS 391	DYS 385a***	DYS 385b***	DYS 42		
—	—	—	—	10	—	—	12		
DYS 389-1**	DYS 392	DYS 389-2**	DYS 458	DYS 459a	DYS 459b	DYS 455***	DYS 454***	DYS 447	DYS 437
—	13	—	17	9		11	11	—	—
DYS 438	DYS 531	DYS 578	DYS 395S1a	DYS 395S1b	DYS 590	DYS 537	DYS 641	DYS 472	DYS 406S1
12	12	9	15	16	9	10	10	8	—
DYS 511	DYS 425	DYS 413a	DYS 413b	DYS 557	DYS 594	DYS 436	DYS 490	DYS 534	DYS 450
—	—	23	—	16	10	12	—	16	8
DYS 444	DYS 481	DYS 520	DYS 446	DYS 617	DYS 568	DYS 487	DYS 572	DYS 640	DYS 492
—	22	—	—	12	11	0	—	—	13
DYS 565	DYS 461***	DYS 462	GATA A10	DYS 635	GAAT1B07	DYS 441	DYS 445	DYS 452	DYS 463
12	12	11	0	—	—	—	—	—	—
DYS 434	DYS 435	DYS 485	DYS 494	DYS 495	DYS 505	DYS 522	DYS 533	DYS 549	DYS 556
—	0	16	9	—	—	0	—	12	11
DYS 575	DYS 589	DYS 636	DYS 638	DYS 643	DYS 714	DYS 716	DYS 717	DYS 726	DXYS156-Y
—	—	12	11	—	25	—	—	—	—

Try it yourself: bit.ly/craig_venter_haplotype_updated



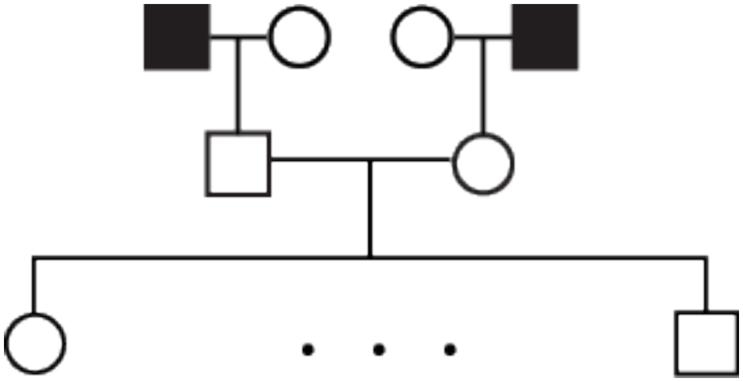
Can we identify anonymous personal genomes?

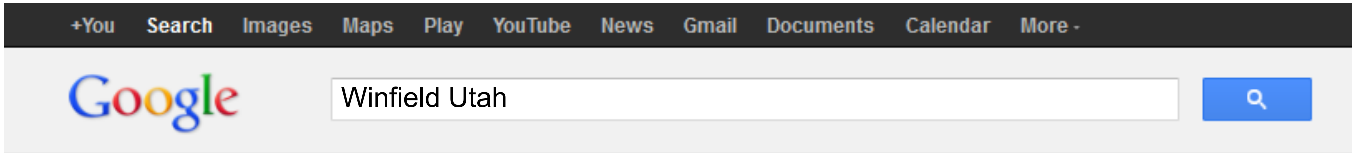
Recovering the identifies of CEU individuals


1000 Genomes
A Deep Catalog of Human Genetic Variation
EMBL-EBI
SORENSON MOLECULAR GENEALOGY FOUNDATION
ysearch

10 CEU genomes

8 Surname predictions with Utah ancestry





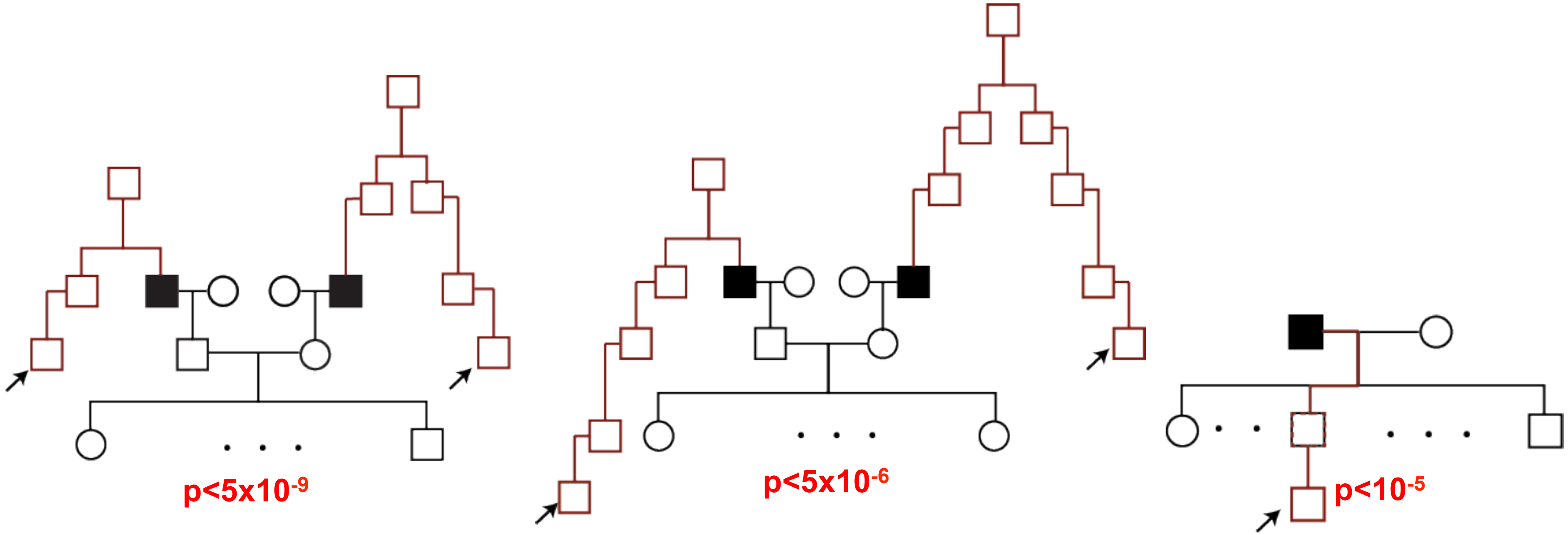
Winfield Utah

Found an obituary that has the exact description of the pedigree

CORIELL INSTITUTE FOR MEDICAL RESEARCH

Probability of a random match < 5×10^{-9}

Beginner’s luck?



■ Successful surname recovery (targeted individual)

↗ Person tested by genetic genealogy service (source)

— Patrilineal line from source to target

Breaching the privacy of close to **50** CEU samples.

Summary

Our approach:

- No experimental work involved.
- The identifying information propagates via deep genealogical ties.
- The attack completely relies on public resources.

Testing close to 1000 Y-STR haplotypes,
demonstrating complete identification of Venter and close to 50 CEU individuals.

IMHO, recommendations

1. Consent:
   - Be honest about risks. Be honest about benefits.

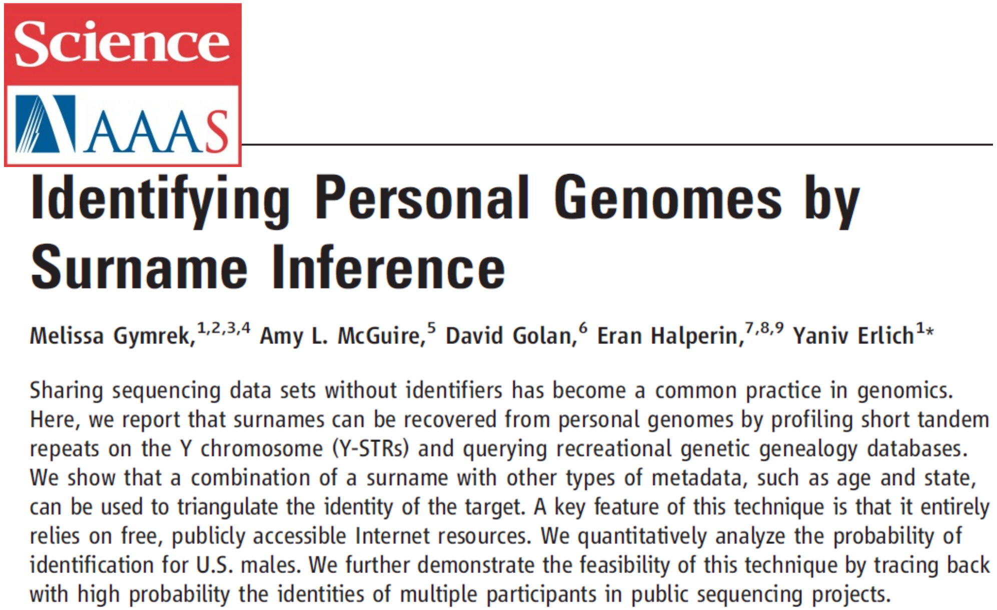
2. Multi-tier approach:
   - Give participants options for data sharing.

3. Proactive approach:
   - Keep mapping risks. Friendly hacking is far better than a real one.

4. Technical solutions:
   - We did not explore those enough. Much more to do here.

Acknowledgements

Melissa Gymrek (HST – Harvard/MIT)
Amy McGuire (Baylor)
David Golan (Tel-Aviv University)
Eran Halperin (Tel-Aviv University)



Science
AAAS

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

Open Access (with FREE registration)

Funding:

Andria and Paul Heafy
Jim and Cathy Stone